ARTÍCULO ORIGINAL

Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas

Data mining techniques aplied to diagnosys of clinical entities

Frank Dávila Hernández,[I] Yovannys Sánchez Corales,[II]

[I]Centro de Informática Médica (CESIM). Departamento Atención Primaria de Salud. Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. E-mail: fdavila@uci.cu
[II]Centro de Informática Médica (CESIM). Departamento Atención Primaria de Salud. Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. E-mail: yscorales@uci.cu

RESUMEN

Disminuir el error médico y mejorar los procesos de salud es prioridad de todo el personal sanitario. En este contexto surgen los "Sistemas Clínicos de Soporte para la Toma de Decisiones" (CDSS), los cuales son un componente fundamental en la informatización de la capa clínica. Con la evolución de las tecnologías gran cantidad de datos han podido ser estudiados y clasificados a partir de la minería de datos. Una de las principales ventajas de la utilización de esta, en los CDSS, ha sido su capacidad de generar nuevos conocimientos. Con este fin se propone, mediante la combinación de dos modelos matemáticos, cómo se puede contribuir al diagnóstico de enfermedades usando técnicas de minería de datos. Para mostrar los modelos utilizados se tomó como caso de estudio la hipertensión arterial. El desarrollo de la investigación se rige por la metodología más utilizada actualmente en los procesos de Descubrimiento de Conocimiento en Bases de Datos: CRISP-DM 1.0, y se apoya en la herramienta de libre distribución WEKA 3.6.2, de gran prestigio entre las utilizadas para el modelado de minería de datos. Como resultados se obtuvieron diversos patrones de comportamiento con relación a los factores de riesgo a sufrir hipertensión mediante técnicas de minería de datos.

Palabras clave: CRISP-DM, hipertensión arterial, KDD, minería de datos, diagnóstico clínico, WEKA.

ABSTRACT

Reduce medical errors and improve health processes is a priority of all health personnel. In this context arise the "Clinical Support Systems for Decision Making" (CDSS), which are a key component in computerization of the clinical layer. With the evolution of technologies, large amounts of data have been studied and classified based on data mining. One of the main advantages of using this in the CDSS, has been its ability to generate new knowledge. For this purpose, this paper presents, by combining two mathematical models, a way to contribute to the diagnosis of diseases using data mining techniques. Hypertension was taken as a case study to show the models used. The research development methodology follows the most used processes of knowledge discovery in databases: CRISP-DM 1.0, and relies on the free distribution tool WEKA 3.6.2. We obtained different patterns of behavior in relation to risk factors for developing hypertension using data mining techniques.

Key words: CRISP-DM, data mining, arterial hypertension, KDD, clinical diagnosys, WEKA.

INTRODUCCIÓN

La Universidad de las Ciencias Informáticas (UCI), posee varios centros de desarrollo de software. El Centro de Informática Médica (CESIM) es uno de ellos, encargado del desarrollo de aplicaciones para el sector de la salud; entre estas se encuentra el Sistema Integral para la Atención Primaria de la Salud (alas SIAPS), el cual posee un componente de tipo Sistema Clínico de Soporte para la Toma de Decisiones (CDSS),[1] para que facilite el procesamiento analítico en línea y la minería de datos y que servirá además al resto de los ambientes bajo un escenario tecnológicamente sólido. Actualmente en el Centro de Toma de Decisiones se está manejando la información con técnicas estadísticas; sin embargo, con estas técnicas no se está aprovechando al máximo la información almacenada.

Las Historias Clínicas Electrónicas (HCE)[2] pertenecientes al alas SIAPS, se encuentran almacenadas en un gran repositorio y su información se envía periódicamente a un Datamart.[3] Dado el gran volumen de datos acumulado en él, y la incapacidad de los especialistas de identificar patrones de comportamiento y extraer conocimiento oculto en los datos almacenados para apoyar sus decisiones, surge la necesidad de aplicar la minería de datos.

En la actualidad, la Hipertensión Arterial se ha convertido en una de las primeras causas de muertes en el mundo. Según el reporte de la Organización Mundial de la Salud (OMS) del 2012[4] 1 de cada 3 personas en el mundo padece de Hipertensión Arterial; además agrega que 1 de cada 10 personas es diabética. Algunos autores como Cumbá,[5] coinciden que anualmente existen 7.2 millones de muertes por enfermedades del corazón. La hipertensión arterial es la segunda causa de muerte a nivel mundial, se reconoce internacionalmente como "muerte silenciosa" pues en la mayoría de los casos los pacientes tienden a ser asintomáticos.

Debido al gran volumen de datos existentes en el datamart, se dificulta la toma de decisiones de los especialistas para realizar un análisis rápido y efectivo y de esta manera encontrar información útil y valiosa oculta en ellos; por otra parte, la no predicción del comportamiento futuro de algunos problemas de salud presentes en las HCE con un alto porcentaje de certeza, basado en el entendimiento del pasado.

La minería de datos[6] es un área de la inteligencia artificial que permite darle solución al problema descrito, la misma se basa en varias disciplinas, algunas de ellas más tradicionales, se distingue de ellas en la orientación más hacia el fin que hacia el medio. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido a la información computarizada que nos rodea hoy en día y que permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones.

En este artículo se propone exponer, mediante la combinación de dos modelos matemáticos, cómo se puede contribuir al diagnóstico de enfermedades, usando técnicas de minería de datos.

MATERIAL Y MÉTODOS

Para mostrar la forma de combinar los modelos, se tomó como caso de estudio la hipertensión arterial. Esta entidad se encuentra con relativa frecuencia en las personas que trabajan y/o estudian en nuestra universidad, lo que permitió disponer de una base de datos propia que sirviera de ejemplo, a pesar de ser una base pequeña y de personas relativamente jóvenes. La hipertensión arterial no es la entidad más apropiada como ejemplo para el uso de la minería de datos, ya que está bastante bien estudiada y no necesita someterse a estas técnicas informáticas modernas para establecer su diagnóstico positivo. Por este motivo, es importante destacar que los datos analizados en este trabajo pudieran no corresponderse con la realidad. Sin embargo debe prestarse mayor atención a la importancia que tiene la utilización de dichas técnicas en la informática aplicada a la medicina.

Metodología computacional, tecnologías y lenguajes utilizados

Cuando se va a realizar un proyecto de minería siempre es necesario contar con una metodología que guíe todo el proceso. En este caso, se seleccionó CRISP-DM versión 1.0 como metodología de desarrollo a utilizar en el proceso de Minería de Datos.

La CRISP-DM (Cross Industry Standard Process for Data Mining) es una metodología de libre distribución que puede trabajar con cualquier herramienta para desarrollar cualquier proyecto. Esta metodología estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. Fue diseñada de forma neutra a la herramienta que se utilice para el desarrollo del proyecto, es de distribución libre y se encuentra en constante perfeccionamiento por parte de la comunidad internacional.[7]

Para realizar el pre-procesado, los que deseen extraer conocimientos a partir de datos deben apoyarse en herramientas de software que les faciliten la tarea. Después de haber realizado un análisis exhaustivo y un estudio comparativo entre aquellas que gozan de mayor popularidad en el mercado se selecciona WEKA versión 3.6.2 como herramienta a utilizar en el proceso de minería de datos.

WEKA (Waikato Environment for Knowledge Analysis)[8] es una herramienta visual de libre distribución bajo licencia GNU desarrollada por un equipo de investigadores de la Universidad de Waikato de Nueva Zelanda. La herramienta está implementada en Java. Es interesante remarcar que, dado que se trata de una herramienta bajo licencia GNU, es posible actualizar su código fuente para incorporar nuevas utilidades o modificar las ya existentes, de ahí que podamos encontrar toda una serie de proyectos asociados a WEKA que permiten garantizar la continua evolución y adaptación de dicha herramienta.[9]

PostgreSQL es un sistema de gestión de base de datos (SGBD) objeto-relacional que posee una gran escalabilidad. Es capaz de ajustarse al número de computadoras y a la cantidad de memoria que posee el sistema de forma óptima, pudiendo soportar una mayor cantidad de peticiones simultáneas de manera correcta. Es multiplataforma, se seleccionó teniendo en cuenta la necesidad de utilizar herramientas libres, para el desarrollo, además de que es un gestor confiable, estable, con control de concurrencia y funcionalidades que lo destacan como uno de los SGBD más potentes en la actualidad.[10]

Trabajos relacionados

Actualmente el panorama es alentador con respecto al desarrollo de aplicaciones que utilizan la minería de datos. Existen un conjunto de técnicas y herramientas capaces de ayudar a la toma de decisiones de los expertos. A pesar de ser relativamente joven, la minería de datos presenta aplicaciones en casi todos los sectores de la sociedad. En la salud, a nivel internacional se destaca la "Aplicación de técnicas de minería de datos para el diagnóstico prematuro del cáncer de mamas". Este sistema se encarga de realizar un diagnóstico del cáncer de mama a partir de una base de datos de imágenes de mamografías.[11] En Cuba se han desarrollado investigaciones como por ejemplo "Aplicaciones de la minería de datos para el análisis de la Información Clínica". Este estudio se basa en el apoyo a la toma de decisiones a partir de coronariografías realizadas a pacientes que padecen cardiopatías isquémicas.[12] La UCI tampoco ha estado ajena al desarrollo de aplicaciones que emplean la minería de datos, y en ese sentido se destaca el "Diagnóstico de enfermedades de transmisión sexual mediante técnicas de inteligencia artificial", que utiliza la información proveniente de un documento Excel para la creación de una aplicación basada en reglas que ayuda a diagnosticar si una persona está infectada de blenorragia o clamidia.[13]

En los tres casos anteriormente mencionados existe una limitante común si se compara con el sistema que este trabajo propone, y es que la información que utilizan para generar los modelos proviene de diversas fuentes y en distintos formatos y no permiten extraerla a partir de un Repositorio Centralizado de Documentos Clínicos.

Algoritmos utilizados

Para el desarrollo de la investigación se seleccionaron dos algoritmos, el J48 dentro de la técnica supervisada Árboles de Decisión y el Simple K-Means para el desarrollo de la técnica no supervisada Agrupamiento. Los mismos fueron seleccionados debido a que son, de acuerdo a la bibliografía consultada, los más utilizados mundialmente dentro de las técnicas a la que pertenecen. El algoritmo J48 amplía las funcionalidades del C4.5, tales como permitir la realización del proceso de post-poda del árbol mediante un método basado en la reducción del error o que las divisiones sobre las variables discretas sean siempre binarias. Este algoritmo permite modelar el resultado del árbol de decisión en lenguaje SQL, tiene una gran velocidad computacional y existe una acertada fiabilidad de los resultados.[14] El algoritmo Simple K-Means pertenece al grupo de algoritmos de partición-optimización, garantiza una elevada semejanza intra-clúster y desemejanza inter-clúster. Este algoritmo presenta como propiedades fundamentales gran velocidad, la cual puede ser considerable cuando se trata de grandes volúmenes de datos, devuelve al usuario buenos resultados y da la posibilidad de cambiar los puntos iníciales y obtener resultados diferentes.[15]

Solución

Lo primero será crear una vista minable, para ello se deben realizar con anterioridad, según CRISP-DM, varios pasos que posibilitarán la adecuada configuración de los registros que se desean analizar. Los mismos se describen brevemente a continuación.

Para recolectar los datos necesarios en la investigación se hizo un análisis de la hipertensión arterial, para lo cual se le realizaron encuestas a los especialistas en este tema. Cada una de las variables que se tuvieron en cuenta fue localizada en las tablas del almacén de HCE y posteriormente descritas para optimizar la comprensión de las mismas. Los datos contenidos en el almacén fueron sometidos a un riguroso análisis basado fundamentalmente en cuanto a representación de la realidad, consistencia, campos innecesarios, campos vacíos y datos de naturaleza híbrida o poco genuina.

Una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. En este punto se deciden seleccionar los atributos y tuplas que serán incluidos en el proceso de minería.

Atributos: genero_paciente, etnia_paciente, edad_paciente.

Tuplas: Antecedentes familiares de enfermedades cardiovasculares, de diabetes mellitus, de hipertensión arterial y de enfermedades renales; antecedentes personales de enfermedades endocrinas, de enfermedades cardiovasculares y de enfermedades renales; disnea, edemas, palpitaciones, náuseas, cefalea y dolor abdominal.

Posteriormente se analizan los datos que son necesarios para el proyecto y se combinan con el objetivo de obtener la información que proviene de las diferentes dimensiones del almacén de datos integradas en una sola tabla: *pre_vista_minable*.

A esta tabla se le aplicaron un conjunto de transformaciones para las cuales se hizo necesaria la creación de un software desarrollado en el IDE NetBeans, el cual funciona como intermediario entre la tabla *md.pre_vista_minable* y *md.vista_minable_j48* (Fig. 1), tabla que almacena los datos que serán utilizados para la creación del modelo mediante árboles de decisión. Se generó además de la tabla *md.vista_minable_j48*, una llamada *md.vista_minable_skm* (Fig. 2), tabla que almacena los datos que serán utilizados para la creación del modelo mediante agrupamiento, la misma es un duplicado de *md.vista_minable_j48*, la diferencia radica en que sus tuplas son numéricas, esto permite una mejor asignación de las variables a la hora de calcular los centroides de los grupos.

id_paciente	rango_edad	ge	etnia_pa	AF	AF	AF	AF	AF	AF	AF	CE	DIS	PA	ED	DA
2	Rango2	M	Mestiza	No	No	No	No	No	Si	No	Si	No	No	No	No
3	Rango2	F	Mestiza	No	No	No	No	No	No	No	Si	No	No	No	No
4	Rango1	M	Negra	No	No	No	Si	No	No	No	Si	No	No	No	No
5	Rango2	M	Negra	No	No	No	Si	No	No	No	Si	No	No	No	No
6	Rango2	M	Blanca	No	No	No	No	No	No	No	Si	No	No	No	No
7	Rango1	F	Blanca	No	No	No	Si	No	No	No	Si	No	No	No	No
8	Rango1	F	Mestiza	No	No	No	No	No	No	Si	Si	No	No	No	No
9	Rango2	M	Blanca	No	No	No	No	No	No	No	No	No	No	No	No
10	Rango2	M	Blanca	No	No	No	No	No	No	No	Si	No	No	No	No
11	Rango3	F	Negra	No	No	No	Si	No	No	No	No	No	No	No	No
12	Rango3	M	Negra	No	No	No	No	No	No	Si	Si	No	No	No	No

Fig. 1. Fragmento de la tabla *vista_minable_j48*.

id_pacier	rango_ed	ge	etnia_pa	AF	AF	AF	AF	AF	AF	AF	CE	DIS	PA	ED	DA
50	2	1	1	0	0	0	0	1	0	0	0	0	0	0	0
32	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
689	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0
254	3	1	1	1	0	0	0	1	0	0	0	0	1	1	0
575	3	1	1	0	1	0	0	0	0	0	0	0	1	0	0
355	3	1	1	0	1	0	0	0	0	0	1	0	0	0	0
384	3	1	1	0	1	0	0	0	0	0	1	0	0	0	0
395	3	1	1	0	1	0	0	0	0	0	1	0	0	0	0
696	3	2	1	0	0	0	1	1	0	1	1	0	0	0	0
574	3	2	1	1	1	1	1	0	0	0	0	1	0	1	0
658	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0
31	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0

Fig. 2. Fragmento de la tabla *vista_minable_skm*.

RESULTADOS Y DISCUSIÓN

Los datos utilizados en esta investigación fueron recopilados de 78 historias clínicas de pacientes hipertensos en la Universidad de las Ciencias Informáticas. Debe insistirse en resaltar que al ser datos de pacientes con características específicas, en su gran mayoría jóvenes, esta no es una muestra representativa de la entidad.

Sin embargo, la investigación se propone analizar relaciones entre los factores de riesgos (antecedentes patológicos tanto personales como familiares, problemas de salud y hábitos personales) y determinar mediante la técnica de árboles de decisión cuáles son los patrones o comportamientos genéricos que caracterizan a los pacientes que acuden a consulta y que permiten ayudar a predecir la enfermedad, y mediante agrupamiento cuáles son los grupos de edades, regiones poblacionales, y otros datos de interés, que más son afectados por la hipertensión arterial; así como establecer relaciones entre las variables analizadas y cómo influyen unas con respecto a las otras.

A continuación se describen los modelos obtenidos así como los patrones identificados para cada uno de ellos, sin ánimo de generalizarlos, brindando algunos detalles que servirán para una mayor comprensión de los mismos. Seguidamente se muestra el modelo obtenido después de haber aplicado el algoritmo Simple K-Means sobre los datos de entrenamiento almacenados en la tabla *vista_minable_skm*. Se procedió a agrupar el set de datos en tres grupos. Para la ejecución de este algoritmo es necesario seleccionar un número, denominado semilla, para realizar una distribución aleatoria inicial a partir de la cual el algoritmo comience las sucesivas iteraciones. Para la selección de este número se realizaron 20 corridas consecutivas probando distintas semillas y se seleccionó aquella que minimizaba la suma del error cuadrático (semilla igual 8). Si bien este método heurístico no garantiza la semilla óptima, asegura una relativamente buena asignación.[15] En la figura 3 se sintetiza un fragmento del resultado obtenido con WEKA tras la ejecución de Simple K-Means con 3 grupos y una semilla de 8. Antes de realizar un análisis a profundidad sobre este modelo, primero es necesario observar las características de cada grupo obtenido una vez aplicado el algoritmo.

Attribute	Full Data (676)	0 (270)	1 (175)	2 (231)
rango_edad	1	2	3	1
genero_paciente	1	1	2	1
etnia_paciente	2	3	1	2
APP_ER	0	0	0	0
APP_EE	0	0	0	0
APP_EC	0	0	0	0
APF_HTA	0	0	0	1
APF_EC	0	0	0	0
APF_ER	0	0	0	0
APF_DM	0	0	0	0

Fig. 3. Distribución de grupos generado por WEKA aplicando el algoritmo Simple K-Means.

A partir de la interpretación conjunta de los gráficos de dispersión podemos descubrir en el conjunto de datos lo siguiente:

- Grupo 0 (40 %): se destacan las personas que se encuentran entre 45 y 65 años de edad, predomina el sexo masculino y la mayoría de ellos son de raza mestiza. La distribución de los pacientes que tienen antecedentes patológicos familiares de hipertensión arterial es bastante uniforme; sin embargo, se puede apreciar una ligera mayoría de personas que no tienen este tipo de antecedente.

- Grupo 1 (26 %): Muy concentrado por personas de más de 65 años de edad, generalmente del sexo femenino y hay mayor concentración de personas de raza blanca. En este grupo, aunque al igual que en el grupo 0 existe una distribución relativamente uniforme de casos de antecedentes familiares de hipertensión arterial, es más notable que en la generalidad de los casos tampoco presentan antecedentes de esta índole.

- Grupo 2 (34 %): Representa en su mayoría a las personas que son menores de 45 años de edad, generalmente masculinos de raza negra. Se puede apreciar una notable concentración de personas que sí presentan antecedentes de hipertensión arterial en su familia.

- Se puede apreciar que en los 3 grupos la generalidad de los pacientes que se encuentran agrupados son personas que tienen hipertensión arterial.

Una vez analizado el contenido de cada grupo se deducen a grandes rasgos los siguientes patrones:

- En el 40 % de los casos de los pacientes que padecen hipertensión arterial están entre 45 y 65 años de edad, son de sexo masculino y de raza mestiza.

- El 34 % de las personas que padecen hipertensión arterial tienen antecedentes patológicos familiares de la enfermedad y consumen tabaco.

- El 66 % de los casos con hipertensión arterial fueron asintomáticos.

En la figura 4 se muestra un fragmento del árbol obtenido a partir de aplicar el algoritmo J48. Los nodos representan atributos, las ramas representan valores de dichos atributos y los nodos finales representan los valores de la clase. Cada camino del árbol representa una regla.

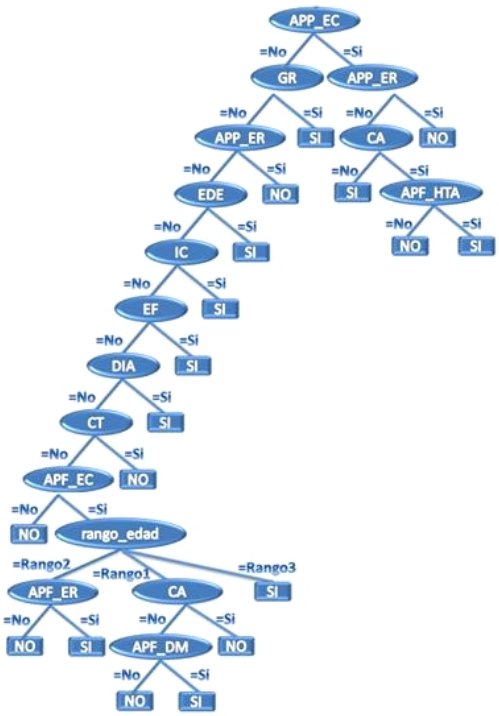

Fig. 4. Fragmento del árbol de decisión generado por WEKA aplicando el algoritmo J48.

CONCLUSIONES

El objetivo fundamental de este trabajo ha sido el estudio y análisis de dos técnicas: clasificación y agrupamiento. A lo largo del mismo, se ha llevado a cabo una importante recopilación bibliográfica y revisión teórica sobre algunos aspectos básicos relacionados con el tema. Se han propuesto además, dos modelos matemáticos cuya combinación puede utilizarse como ayuda al diagnóstico de entidades clínicas. Aunque los algoritmos se propusieron en el sector de la salud, su uso no está restringido a esta área. El primer aporte de este trabajo se centra en la construcción de dos modelos mediante clasificación y agrupamiento, árbol de decisión J48, Simple K-Means respectivamente, con la estrategia de encontrar patrones ocultos en los datos clínicos de pacientes que sufren de hipertensión arterial. Además tendrá un aporte práctico basado en que el Sistema Integral para la Atención Primaria de la Salud contará con un soporte de toma de decisiones que lo convertirá en un sistema más robusto, el mismo permitirá acelerar el proceso de análisis de la información de los especialistas en la toma de decisiones médicas. Finalmente, cabe destacar el hecho de que los dos modelos obtenidos han sido evaluados sobre datos reales, comparando sus resultados con los obtenidos de diferentes procedimientos, mediante las propias evaluaciones de los modelos que ofrece WEKA y el visto bueno de los expertos en HTA.

Esta investigación servirá como base para la realización de otros trabajos de manera que perfeccionen lo descrito anteriormente. Servirá de modelo que puede ser implementado en el CDSS y de esta manera encontrar nuevo conocimiento a partir de otras enfermedades, relacionando sus síntomas, causas y diagnósticos futuros.

REFERENCIAS BIBLIOGRÁFICAS

1. Sanchez, Y. et al. Centro de Toma de Decisiones en el Sistema Integral para la Atención Primaria de Salud. La Habana, Cuba: Memorias del Evento INFORMÁTICA; 2011. ISBN: 978-959-7213-01-7

2. Cosialls, D. Información para la gestión clínica. Contrato de servicio Vol. 2. Madrid: ELSEVIER ESPAÑA; 2000.

3. Kimball R, Ross M. The Data Warehouse Toolkit. Canberra, Australia: John Wiley & Sons Incorporated; 2006. ISBN: 0-471-15337-0.

4. OMS. Estadísticas Sanitarias Mundiales 2012. US: World Health Organization; 2012. ISBN: 978 92 4 356444 9.

5. Fernández Cumbá E. Propuesta didáctica para la promoción de salud en el caso de la hipertensión arterial en los pacientes de la Universidad de las Ciencias Informáticas. La Habana: Instituto Superior Politécnico José A. Echeverría; 2008.

6. Hand D, Mannila H, Smyth P. Principles of Data Mining. Cambridge, Massachusetts London England: Massachusetts Institute of Technology; 2001.

7. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0. Guía paso a paso de minería de datos. [Citado el 12 Ago. 2012]. Disponible en: http://www.crisp-dm.org

8. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann Series in Data Management Systems; 2005.

9. Weka. [homepage] Nueva Zelanda: Universidad de Waikato. [Citado el: 9 de Mayo de 2011]. Disponible en: http://www.cs.waikato.ac.nz/ml/weka/

10. Pecos D. PostGreSQL vs. MySQL. [Citado el 28 de Enero de 2011]. Disponible en: danielpecos.com/docs/mysql_postgres/x15.html

11. Vallejo Delgado N, Rodríguez Jara F. Aplicación de técnicas de minería de datos para el diagnóstico prematuro de cáncer. [citado el 13 Nov. 2012]. Disponible en: http://www.it.uc3m.es/jvillena/irc/descarga.htm?url=practicas/08-09/02.pdf

12. Rosete Suárez A, Rodríguez Díaz A, Acosta Sánchez R. Predicción de pacientes diabéticos. Preprocesado para Minería de Datos. Revista Cubana de Informática Médica. 2009 [Citado el 3 de Nov. 2011]; 9(1). Disponible en: http://www.rcim.sld.cu/revista_18/articulos_htm/prediccionpaciente.htm#t

13. Bañobre Corpas Y, Brossard González Y. Diagnóstico de Enfermedades de Transmisión Sexual mediante técnicas de Inteligencia Artificial. La Habana: Universidad de las Ciencias Informáticas, Facultad 5; 2009.

14. Marante Jacas D, Marante Jacas D. Aplicación de la minería de datos para la exploración y detección de patrones delictivos. La Habana: Universidad de las Ciencias Informáticas, Facultad 8; 2008.

15. Perversi I. Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina. [Citado el: 9 de Noviembre de 2011]. Disponible en: http://laboratorios.fi.uba.ar/lsi/rgm/tesistas/PERVERSI-tesisdegradoeningenieria.pdf

Recibido: 12 de octubre de 2012.
Aprobado: 13 de noviembre de 2012.